Simplified inverse filter tracked affective acoustic signals classification incorporating deep convolutional neural networks

Simplified inverse filter tracked affective acoustic signals classification incorporating deep convolutional neural networks

Yuxiang Kuang[1], Qun Wu[2 *], Ying Wang[3 *], Nilanjan Dey[4], Fuqian Shi[5], Rubén González Crespo [6], and R. Simon Sherratt[7]

1. Arts College, Jiangxi University of Finance and Economics, Nanchang, PR China
2. Institute of Universal Design, Zhejiang Sci-Tech University, Hangzhou, PR China
3. Department of Industrial Design at College of Art and Design, Zhejiang Sci-Tech University, Hangzhou, 310018, PR China
4. Department of Information Technology, Techno International New Town, West Bengal, India
5. Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ, 08903, USA
6. Department of Computer Science and Technology, Universidad Internacional de La Rioja, Logroño, Spain
7. Department of Biomedical Engineering, the University of Reading, RG6 6AY, UK

***Corresponding authors**:
Qun Wu and Ying Wang, No. 928, the 2nd St., Xiasha High-Tech Education Zone, Hangzhou City, Zhejiang, PR China; E-mail: wuq@zstu.edu.cn and winered@zstu.edu.cn; Tel.: +86-571-86843291.

Abstract

Facial expressions, verbal, behavioral, such as limb movements, and physiological features are vital ways for affective human interactions. Researchers have given machines the ability to recognize affective communication through the above modalities in the past decades. In addition to facial expressions, changes in the level of sound, strength, weakness, and turbulence will also convey affective. Extracting affective feature parameters from the acoustic signals have been widely applied in customer service, education, and the medical field. In this research, an improved AlexNet-based deep convolutional neural network (A-DCNN) is presented for acoustic signal recognition. Firstly, preprocessed on signals using simplified inverse filter tracking (SIFT) and short-time Fourier transform (STFT), Mel frequency Cepstrum (MFCC) and waveform-based segmentation were deployed to create the input for the deep neural network (DNN), which was applied widely in signals preprocess for most neural networks. Secondly, acoustic signals were acquired from the public Ryerson Audio-Visual Database of Affective Speech and Song (RAVDESS) affective speech audio system. Through the acoustic signal preprocessing tools, the basic features of the kind of sound signals were calculated and extracted. The proposed DNN based on improved AlexNet has a 95.88% accuracy on classifying eight affective of acoustic signals. By comparing with some linear classifications, such as decision table (DT) and Bayesian inference (BI) and other deep neural networks, such as AlexNet+SVM, recurrent convolutional neural network (R-CNN), etc., the proposed method achieves high effectiveness on the accuracy (A), sensitivity (S1), positive predictive (PP), and f1-score (F1). Acoustic signals affective recognition and classification can be potentially applied in industrial product design through measuring consumers' affective responses to products; by collecting relevant affective sound data to understand the popularity of the product, and furthermore, to improve the product design and increase the market responsiveness.

Keywords: AlexNet; deep convolutional neural network; acoustic signals; affective computing; short time Fourier transform

1. Introduction

Affective computing refers to the ability to detect, classify, organize, and respond to human affective communication which can help users get efficient and friendly feelings. This kind of development can also be used by special credits to help people understand the affective world of themselves and others. The research of affective computing makes the formal machine more visual, which is the prerequisite for realizing

human-computer interaction [1]. To date, affective information extraction based on acoustic features is widely used in language, affective, and affective recognition, and making much more critical issues in the field of intelligent computing. The sound itself has no affective, but auditory signals (such as loudness and fluency), lip muscle tension and facial expressions during pronunciation can give a psychological or innate feeling to people [2, 3]. Sound is an analog signal, and the time domain waveform of the sound only represents the relationship of the signals' pressure with time and cannot be represented as the features of the sound as well [4]. Therefore, the sound waveform need to be converted into acoustic feature vectors formally; Turner et al. developed a flexibility algorithm to the decomposition process of discrete wavelet transform (DWT), so-called wavelet packet transform (WPT) for speaker identification [5]; Yuan, X.-C at el. [6] also used wavelet packet analysis for speaker-independent affective recognition and improved the recognition rates in EMODB (Berlin Database of Affective Speech) and EESDB (an elderly affective speech database). At present, there are many acoustic feature extracting methods (FEM), such as Mel frequency Cepstrum coefficient (MFCC), linear prediction cepstral coefficient (LPCC), multimedia content description interface (MPEG7), etc. [7]. Among them, MFCC is based on Cepstrum, which is more in line with the principle of human hearing and is therefore the most commonly in the most effective acoustic feature extraction algorithms. Before extracting the MFCC, the acoustic signals need to be pre-processed, including analog/digital (AD) transferring, pre-emphasizing, and windowing [8].

The classification of acoustic affective features is mainly subdivided into three categories including prosodic features (super-segment features / hyper-linguistic features) including duration-related features, fundamental frequency-related features, and energy-related features [9]; and sound quality features, such as spectrum-based correlation analysis feature is a reflection of the correlation between the change in the form of the vocal tract and the vocalization motion [10 11]. Acoustic feature extraction is the important step in the whole acoustic recognitions process. It includes the fundamental frequency characteristics, such as pitch, it is the reciprocal of the vocal cord vibration frequency, which refers to the period when a person makes voiced sounds, and the airflow causes the vocal cords to vibrate through the soundtrack. The period of vocal cord vibration is the pitch period. The fundamental frequency contains a large number of features that characterize speech affective, which are crucial in acoustic affective recognition. The change range is 50-500Hz; and the detection difficulty is relatively high [12, 13]. Commonly, fundamental frequency feature extraction methods are used for the autocorrelation function (ACF)-time domain, average amplitude difference method (AMFD)-time domain and wavelet method (WM)-frequency domain; C.K, Y., et al selected higher order spectral features in a set for affective recognition by using 28 bi-spectral features and 22 bi-coherence features [14].

In other way, the formant characteristics also can be used for acoustic recognition; from an acoustic point of view, it means that the sound channel can be regarded as a sound tube with a non-uniform cross-section, while the frequency of the sound excitation signal matches the frequency of the sound channel; and the sound channel will resonate. The resulting waveform is called a formant. Formants are one of the most important parameters of speech signal processing, which determines the sound quality in vowels. Its parameters include formant frequency and formant bandwidth. Daneshfar et al. [15] introduced a high-dimensional hybrid feature vector for dealing with spectral-prosodic features of speech and performed high effectiveness in recognize EMODB. The positions of the formants of different affective pronunciations are different. When the affective state changes, the peak values of the first three formants change greatly, and the peaks are the first, second, and the third formant from low to high. Generally, the average value, maximum value, minimum value, dynamic change range, average change rate, mean square deviation of the first, second, and third formant are selected [16, 17].

Before deep learning is applied, spectrogram, MFCC are commonly used as for acoustic signal processing; while in other hand, the Koff model is still a common method for acoustic affective recognition; and traditionally, convolutional neural networks (CNN) and time-domain pyramid matching also were used to extract and recognize affective features in acoustic signals [18, 19]. As deep-learning-based recognition methods are developed recently, unsupervised learning program-a feature detector layer may be created without labeling data; and a reconstruction learning goal is used to "pre-training" several layers of increasing complexity feature detector. The first important application of this pre-training method is in acoustic feature recognition [20-22]. This method also has been used to calculate a series of probability values corresponding to the window of short-term coefficients extracted from a sound sample. These probability values reflect the probability that each segment of speech is represented by a frame in the window. In the standard acoustic

recognition test of the small vocabulary, the training effect of this method broke the record, and soon it is developed to break the standard voice test record of the large vocabulary [23]. The methods of acoustic feature extraction based on deep learning also include some common feature de-dimensionality algorithms, such as principal component analysis (PCA), linear discriminant analysis (LDA), local preservation projection (LPP), multidimensional scale analysis (MDS), Isometric mapping (ISOMAP), local linear embedding (LLE), and Laplacian eigenmaps (LE) [24,26]. The traditional CNN model has a shallow structure and limited ability to model acoustic features. If a CNN and a deep confidence network (DBN) are used to generate a deep convolutional network, a neural network can be added. The depth of the model to enhance the modeling ability of the model will be created [27, 28]. The deep CNN model may improve the accuracy in acoustic recognition task. The proposed residual / highway network allows us to train the neural network deeper [29, 30]. In the process of trying deep CNN, there are roughly two strategies: one is the acoustic model based on the deep CNN structure in the HMM framework. The CNN can be a CNN network structure connected by VGG or residual, or a CLDNN structure. The other is a very popular end-to-end structure in the past two years, such as the use of CNN or CLDNN in the CTC framework to achieve end-to-end modeling, or coarse-grained modeling unit technology, such as low frame rate and chain (LFRC) model. Google attempts to deep CNN's path mainly using a variety of methods and model fusion, such as network-in-network (NiN), batch normalization (BN), and convolutional LSTM (ConvLSTM) [31,32]. However, studies have shown that convolutional neural networks help the training set or tasks with small data differences the most. For most other tasks, the relative word error rate declines generally only 2% up to 3% [33]. As a deep neural network, the AlexNet network structure model proposed by Alex in 2012 detonated the application of neural networks, has been widely used in acoustic affective recognition [34-36]. Furthermore. Boddapati et al. [37] used CNN and DBN for image recognition, such as, AlexNet and GoogLeNet for classifying environmental acoustic signals, and Boloukian and Safi-Esfahani [38] developed an autoencoder neural Turing machine (DN-AE-NTM) model for classifying speech-impaired people.

In this paper, acoustic feature signals are used for speech affective-based calculation and recognition; Music Analysis, Retrieval and Synthesis for Audio Signals (MARSYAS, http://marsyas.info), an audio signals analysis, retrieval and synthesis tools for basic features calculation. By separating preprocessing and waveform-based processing, an improved AlexNet deep convolutional neural network (ADCNN) is finally used to classify acoustic affective dataset from public datasets. The framework of the research is shown in Fig.1. The rest of the paper structure is organized as follows; Section 2 introduces the data preprocessing and feature extraction; Section 3 presents the modeling and method; Section 4 presents the results and discussion, and Section 5 presents the conclusion.

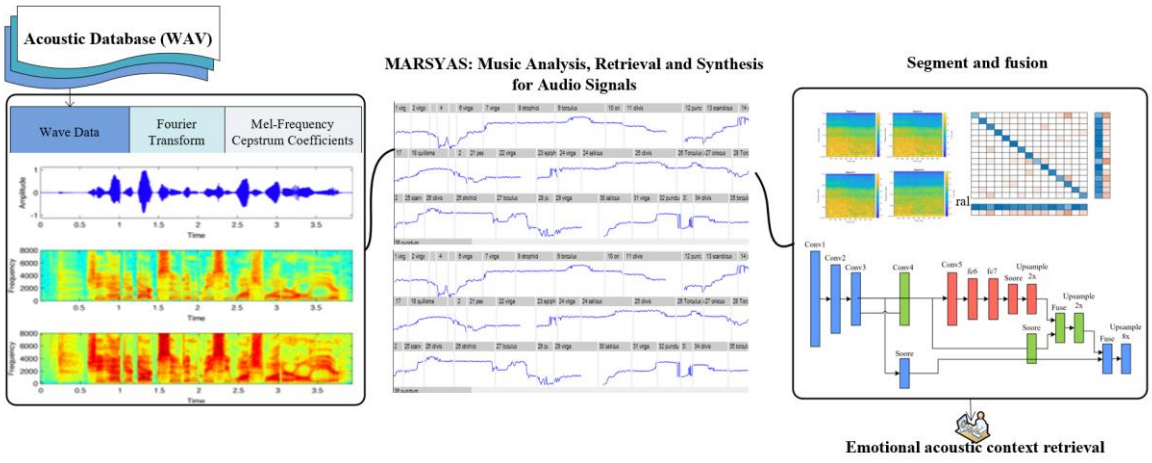


Figure 1. the framework of affective acoustic recognition using an improved AlexNet-based deep neural networks

2. Modelling
2.1 Preprocessing

The purpose of acoustic signals preprocessing is to highlight the useful information contained in the speech acoustic signals as much as possible and reduce the impact of other useless information (noisy). The following sections briefly introduce the sampling quantization, windowing and framing, and endpoint detection.

2.1.1 Sampling and Quantization

To perform post-framing windowing and endpoint detection processing, the voice signal needs to be converted from a continuous analog signal to a discrete digital signal. The digitization of speech signals includes two processes: sampling and quantizing. The time interval in the sampling process is determined by the sampling frequency. The larger the sampling frequency, the smaller the time interval and the better the quality of the resulting speech acoustic signal. However, the sampling frequency is not as large as possible. Due to the limited hearing ability of the human, while the sampling frequency exceeds a certain range, so the acoustic signal will contain more redundant information that is not related to affective expression, which may affect human affective. In the field of speech acoustic processing, the sampling frequencies of 11.25kHz, 22.05kHz and 44.1kHz respectively represent the three qualities of low, medium and high signals. For the research of speech affective recognition, generally adopting medium-quality speech signals can satisfy the demand. So, a sampling frequency of 22.05kHz can be used by sampling. After sampling the original acoustic signal, only discrete digitalization is realized in time, and the amplitude needs to be digitized by quantization. The most important parameter in the quantization process is the number of quantization bits and 8-bit, 12-bit, and 16-bit are commonly used. The higher the number of bits, the higher the accuracy, and the greater the storage space required as well.

2.1.2 Windowing and Framing

Although the voice signal is a non-stationary signal that changes with time, it also has short-term stability, that is, within a short time range (usually considered to be 10-30ms), the voice signal remains basically unchanged. Therefore, in order to analyze the characteristic parameters of the voice signal, the voice signal must be segmented first, which is commonly referred to as framing. It is determined by two parameters, frame length and frame shift. Frame length refers to the length of each frame. Frame shift indicates the degree of overlap between two adjacent frames. In this paper, when the affective speech is framed, the frame length is 512 sampling points, and the frame shift is 256 sampling points. The framing process usually requires the use of a fixed-length window moving in one direction. This is the windowing operation mentioned. The relationship between the digitized speech signal $x(n)$ and the window function $w(n)$ can be expressed as

$$x_w(n) = x(n) \times w(n) \tag{1}$$

where, the $w(n)$ has the form of rectangular windows, Hamming windows and Hann windows are shown as follows,

$$w(n) = \begin{cases} 1 & 0 < n \le N-1 \\ 0 & others \end{cases} \tag{2}$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\dfrac{2n\pi}{N-1}) & 0 < n \le N-1 \\ 0 & others \end{cases} \tag{3}$$

$$w(n)=\begin{cases}(1-\cos(\frac{2n\pi}{N-1}))/2 & 0<n\le N-1\\ 0 & others\end{cases} \tag{4}$$

where, N is the length of the frame. The main lobe of the rectangular window is narrow, but the peak of the side lobe is high, which causes serious spectrum leakage; the main lobe of the Hamming window is wider, which is about twice that of the rectangular window, but the side lobe attenuation is large, which can effectively reduce the spectrum leakage To reflect the spectral characteristics of short-term signals to a large extent. Hann window and Hamming window are both cosine windows, but the weighting coefficients are different, but they are better than rectangular windows. In this paper, the window function used for framing is the Hamming window.

2.1.3 Endpoint Detection

Endpoint detection is a key step in the speech processing process. The purpose of endpoint detection for voice signals is to find the start and end points of a segment of speech and obtain valid segments of speech. This paper adopts a dual-limit gate endpoint detection method based on short-term energy and short-time zero-crossing rate. The specific steps are as follows:

Step 1: First calculate the energy of each frame of voice signal after framing by,

$$E_n=\sum_{m=0}^{N-1}x_n^2(m) \tag{5}$$

Where, n refers to the n-th frame speech signal, m refers to the m-th sampling point, and N is the frame length.

Step 2: calculate the zero-crossing rate (z_{cr}) of each frame,

$$z_{cr}=\sum_{i=1}^{N}|\,sgn(x_n(i)-\mathrm{sgn}(x_n(i-1))\,| \tag{6}$$

where, $|.|$ means absolute and sgn(.) is a sign function as defined by,

$$\mathrm{sgn}(x)=\begin{cases}-1 & if\ x<0\\ 0 & if\ x=0\\ 1 & if\ x>0\end{cases} \tag{7}$$

Step 3: calculate the average "e" and standard deviation "σ_e" of the short-term energy of the entire voice on a frame basis, and set two thresholds for the initial detection of the voice start and end points, high threshold -15 dB, low threshold is $e+3\sigma_e$ dB. The high threshold is used to determine the starting point of the voice, and the low threshold is used to determine whether the voice ends.

Step 4: calculate the mean value "mv" and standard deviation "σ_{mv}" of the short-term zero-crossing rate of the entire voice on a frame basis, and set the zero-crossing rate threshold as $IZCT=mv+3\sigma_{mv}$. The consonant position at the end completes the second detection of the starting point and ending point of the voice.

2.2 Deduction and Segmentation

2.2.1 features extraction for acoustic signals

The Fourier transform is to de-compose signals into countless sine wave (or cosine wave) signals. Because of its small amount of calculation, Fast Fourier Transform (FFT) has been widely used in the field of signal processing technology, for example, analysis and Synthesis, and multiplex conversion of fully digital time division and frequency division (TDM/FDM) in communication systems, as well as signal filtering and related analysis in the frequency domain; in addition, through radar, sonar, and vibration signals. In order to improve the resolution of the target search and tracking, the FFT is used to analyze the spectrum. The emergence of FFT has played an important role in the development of digital signal processing. In this study, we used FFT to process and reduce the dimension of the speech signal to facilitate the next step of processing. Another deduction method applied in this paper is the simplified inverse filter tracking (SIFT), which is a new version of related processing methods for extracting basic audio signals. The basic idea of this method is to first perform LPC analysis and inverse filtering on the acoustic signal to obtain the predicted residual of the speech signal, then filter the residual signal through an autocorrelation filter, and then perform peak detection. Flatten the obtained spectrum [39]. The Linear Predictive Analysis Method (LPC) is that the acoustic signal can be approximated by a linear combination of several acoustic sampling points in the past. By minimizing the variances between the predicted sample values and the actual output values, a set of linear prediction coefficients can be obtained from the transfer function of the channel. The power spectrum of the channel transfer function can be obtained by modulo H (z), and the bandwidth and center frequency can be detected more accurately based on the power spectrum.

The one of the rest processes in this stage is formant extraction including Cepstrum, which uses the homomorphic unwinding technique to separate the pitch information from the channel information, so that the formant parameters can be directly obtained. This method is more accurate than directly performing the Discrete Fourier Transform (DFT) calculation to obtain the formant, which avoids the fundamental harmonic wave frequency error. The second process used in this research is Mel frequency Cepstrum coefficient (MFCC) extraction. MFCC is a characteristic parameter discovered according to human hearing mechanism, which has a non-linear correspondence with frequency. Below 1000 Hz, the human ear's ability to perceive sound has a linear relationship with frequency, while above 1000 Hz, the human ear's ability to perceive sound has a nonlinear relationship with frequency. MFCC utilizes this non-linear relationship to obtain the spectral characteristics. It is based on the hearing merits of the human ear and is a robust frequency domain speech feature parameter. The human ear uses Mel to subjectively measure the pitch. The pitch of a speech signal of 1000 Hz and 40 dB is specified as 1000 Mel. On the Mel scale, the subjective perception of the human ear on the pitch of speech is linear. The human ear basement membrane is equivalent to a non-uniform filter bank. The cell membranes in different places have different responses to frequency. Each part corresponds to a filter group, and each filter group corresponds to a center frequency and bandwidth. The bandwidth is about 100 Mel [40,41].

Fourier transform the original acoustic signal to obtain the frequency spectrum by using

$$X[k] = H[k]E[k] \tag{8}$$

And the amplitude is presented by

$$\| X[k] \| = \| H[k] \| \| E[k] \| \tag{9}$$

Where, $\| . \|$ is a norm of a vector. Continuously, taking logarithms on both sides, we have that,

$$\log \| X[k] \| = \log \| H[k] \| E[k] \| = \log \| H[k] \| + \log \| E[k] \| \tag{10}$$

And then, taking inverse Fourier transform on both sides, we have that,

$$x[k] = h[k] + e[k] \tag{11}$$

Summarized that Cepstrum is a spectrum obtained by inverse Fourier transform of the Fourier transform of a signal after logarithm operation.

As the human's hearing system is a particular non-linear system, and its sensitivity for different frequency signals is different obviously. So, in the extraction of acoustic features, the human hearing can not only extract semantic information, but also extract the speaker's personal features, which are beyond the reach of

existing acoustic recognition systems. If the characteristics of human auditory perception processing can be simulated in the acoustic recognition system, it is possible to improve the speech recognition rate. MFCC here simplices the human hearing merits; and the linear spectrum is first mapped to the MFCC's nonlinear spectrum based on auditory perception, and then being converted to the Cepstrum. The formula for converting ordinary frequency to Mel frequency is,

$$mel(f) = 2595 * \lg(1 + f / 700) \tag{12}$$

Then, we pass the spectrum through a set of Mel filters to get the Mel spectrum. The formula is,

$$\log X[k] = \log(mel - spectrum) \tag{13}$$

The Cepstrum based analysis is illustrated in Fig. 2. The algorithm for MFCC based acoustic signals processing as described in Algorithm 1.

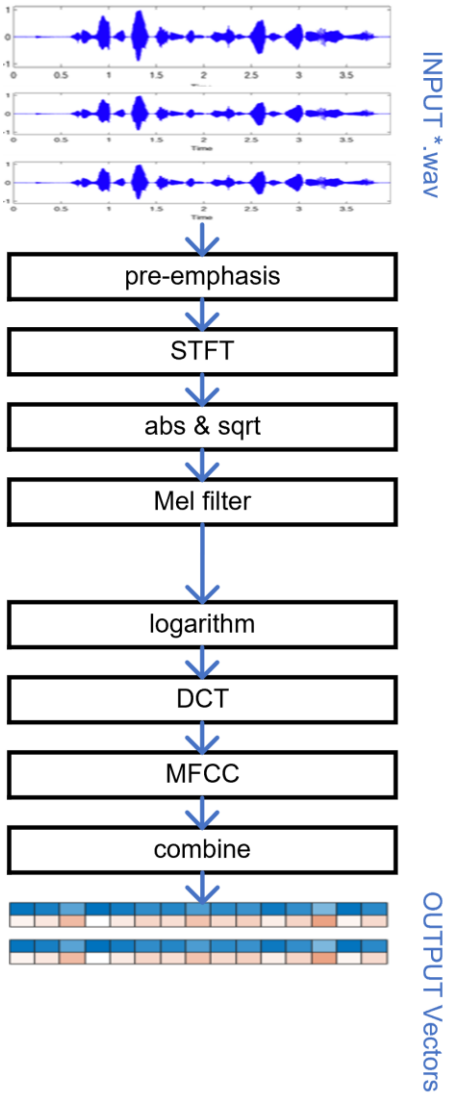


Figure 2. MFCC based acoustic feature extraction process

Algorithm 1: MFCC for acoustic signals processing

REQUIRED: file_wav; Mel filter=26; len_FFT=512; sample_frq=16000Hz; lifter=22; pre-emphasis=0.98; hamming_wind=TRUE; len_frame=400; _shift_frame=160;

OUTPUT: m; ac ();

```
# required wav file
wavdata←audioread(*.wav)
#Mel filter
bank←melbankm(Mel_filt,len_FFT,sample_frq,Hammin_wind)
# normalization
bank←lognormalization(bank)
# calculate dct coefficients
FOR k←1 TO 12
    N←1:26;
    dctcoef(k)←COS((2*n-1) *k*PI/ (2*26)); //PI=3.14159
ENDFOR
# pre-emphasis
y.p←filter(1, pre-emphasis)
# framing
y.p=enframe(y.p,len_frame,shift_frame);
# calculate each frame's MFCC
FOR i←1 TO size (y.p,1)
     y←y.p(i,:);
     y←hamming_wind(y,400);
     y←abs(fft(y,512)); //FFT
     y←y^2;
     c1←dctcoef*log (bank*y (1:257)) ;// if N mod 2==0 then we have that N/2+1=257
     c2←c1*w;
     c2←c2*sqrt (2.0/26);
     m(i)←c2;
ENDFOR
#calculate the first order difference coefficient
dtm=zeros(size(m));
FOR i←3 TO size(m,1)-2
    dtm(i)=-2*m(i-2)-m(i-1) +m(i+1) +2*m(i+2)
ENDFOR
dtm=dtm/3;
#calculate the second order difference coefficient
dtmm=zeros(size(dtm));
FOR i←3 TO size(dtm,1)-2
    dtmm(i)=-2*dtm(i-2)-dtm(i-1) +dtm(i+1) +2*dtm(i+2);
ENDFOR
dtmm=dtmm/3;
#combine two MFCCs
ac←[m dtm dtmm];
RETURN:
     ac, m
```

2.2.2 Time Frequency Separating

Acoustic signal segmentation is a very important issue for signal affective recognition. The automatic accurate positioning and segmentation of continuous acoustic signals is very difficult. It will lead to serious accumulation of neural network training errors in the later period if the signal segmentation is inaccurate; and the segmentation method based on the time domain belongs to the non-model method can be used to optimize the analysis process. Before that, a tool called MARSYAS (Music Analysis, Retrieval and Synthesis for Audio Signals) was used for acoustic signals' features calculation, which is an open source software framework for audio processing with acoustic information retrieval applications. Algorithm 2 introduced the process for time frequency based acoustic signal separating process [42].

Algorithm 2: time frequency separating

REQUIRED: w (wave file)

OUTPUT: wave_data

```
START
  w←readfile(*. wave)
   path← wave_read (path):
    wavfile ← w.open(path,"rb")
    params ← wavefile.getparams()
    # get features using MARSYAS system
    n_channels, sample_width, frame_rate, n_frames ← params [:4]
    datawav ← wavefile.readframes(n_frames)
    wavefile.close()
    wave_data ← np.fromstring(datawav, dtype = np.short)
    if n_channels==1: wave_data.shape←-1,1
    if n_channels==2: wave_data.shape←-1,2
    wave_data ← wave_data.T
    time ← np.arange(0, n_frames) * (1.0/frame_rate)
RETURN
   wave_data[0], time
```

2.3 Classification

2.3.1 AlexNet

AlexNet is developed from LeNet [43] by deepening its network structure and learns much richer and higher dimensional image features. AlexNet has a deeper network structure, uses a cascaded convolutional layer, that is, convolutional layer + convolutional layer + pooling layer to extract the characteristics of the signal, uses Dropout to suppress overfitting, and can use data enhancement technology; use ReLu [44] to replace the previous sigmoid as an activation function, the main structure is as follows:

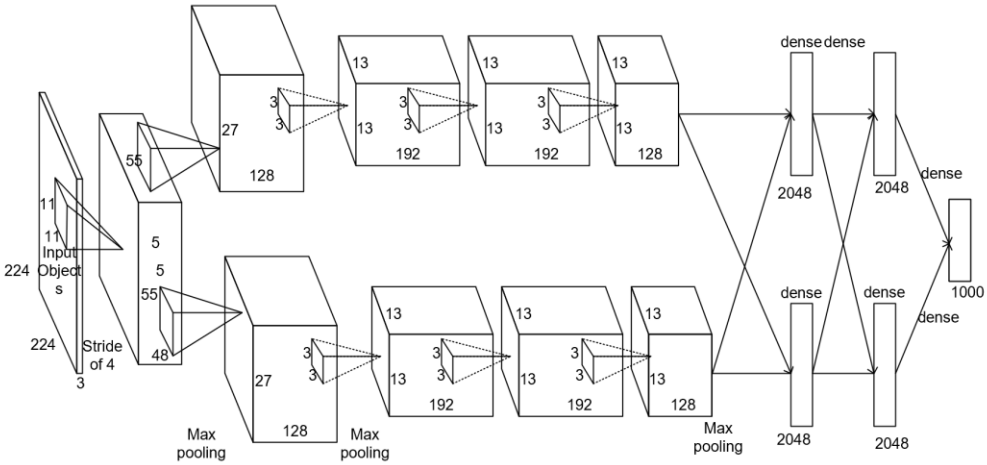


Figure 3. AlexNet based deep neural networks structure

The first layer is convolutional layer 1, and the input is 224 × 224 × 3 images; the number of convolution kernels is 96; in this research, 48 cores is applied, and the size of the core is 11 × 11 × 31; set stride be 4, stride means the step size, and set pad be 0, that means the edge is not extended. The size of the graph after convolution can be calculated as, wide = (224 + 2 * padding-kernel_size) / stride + 1 = 54, height = (224 + 2 * padding-kernel_size) / stride + 1 = 54, and dimension = 96. Then the perform index so called Local Response Normalized (LRN) is followed by pooling using pool_size = (3, 3), stride = 2, and pad = 0, and finally the feature map of the first layer of convolution is acquired. The second layer is convolutional layer 2. The input is the feature map of the previous layer of convolution (first layer); the number of convolutions is 256, and then we have 128 convolution kernels respectively in this research. The size of the convolution kernel is: 5 × 5 × 48; pad = 2, stride = 1; then do the LRN, and finally so max_pooling using pool_size = (3, 3) and stride = 2. The third layer is convolution layer 3. The input is the output of the second layer; the number of convolution kernels is 384, kernel_size = 3 × 3 × 256, padding = 1, third Layer does not do LRN and Pool.The fourth layer is convolution layer 4. The input is the output of the third layer; the number of convolution kernels is 384, kernel_size = 3 × 3, padding = 1; is it the same as the third layer without LRN and Pool. The fifth layer is convolution layer 5. The input is the output of the fourth layer; the number of convolution kernels is 256, kernel_size = 3 × 3, padding = 1. Then, directly perform the max_pooling using pool_size = (3, 3) and stride = 2. The sixth, seventh, and eighth layers are fully connected layers. The number of neurons in each layer is 4096, and the final output softmax is 1000; because as mentioned above, the classification number is 1000. ReLU and dropout are used in the fully connected layer in this research. ReLU nonlinearity (Rectified Linear Unit) is,

$$\tanh(x) = \frac{(e^{x} - e^{-x})}{e^{x} + e^{-x}} \tag{14}$$

The sigmoid function is defined as,

$$f(x) = \frac{1}{1 + e^{x}} \tag{15}$$

In the neural network, the activation function does a non-linear mapping of the output of the neuron, but the range of traditional activation functions such as tanh and sigmoid has a range, and the range of the ReLU activation function has no interval, so the result from ReLU need to be normalized (LRN), the LRN is defined as,

$$b_{(x,y)}^{i} = \frac{a_{(x,y)}^{i}}{(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{(x,y)}^{j})^{2})^{\beta}} \tag{16}$$

Where $a^i_{(x,y)}$ presents the i-th kernel's output in position (x, y) ; n presents the number of the neighbors of $a^i_{(x,y)}$. N is the total counts of kernels; $b^i_{(x,y)}$ is the result of LRN. α, β are parameters.

The final step is overlapping pooling. The general pooling layer does not overlap, so pool_size and stride are generally equal; for example, an image of 8 × 8, if the size of the pooling layer is 2 × 2, then the operation after pooling will get an image with 4 × 4. This setting is called uncovered pooling operation; if stride less than pool_size, then it will generate a covered pooling operation, which is somewhat similar to convolutional operation, so that more accurate result can be acquired. Pooling operations uses overlays in top-1 and top-5 reducing the error rate by 0.4% and 0.3%, respectively. In the process of training the network model, the covered pooling layer is less likely to overfit. So, dropout needs to be carried out for preventing overfitting. In the neural network, dropout is realized through modifying the structure of the neural network itself. For a certain layer of neurons, the neuron is set to 0 by a defined probability. This neuron does not participate in forward and backward propagation, just like in the network. It is deleted in the same way while keeping the number of neurons in the input layer and the output layer unchanged; and then update the parameters according to the learning method of the neural network. In the next iteration, some neurons are randomly deleted again (set to 0) until the end of training. Dropout should be regarded as a great innovation in AlexNet, and now one of the necessary structures in neural networks, dropout can also be regarded as a model combination. The network structure generated each time is different. By combining multiple models, overfitting can be effectively reduced while dropout only requires twice the training time to achieve model combination (similar to average) in much higher efficiencies.

2.3.2 the improved AlexNet-based deep convolutional neural network

The normalization method used in AlexNet is LRN, but LRN has certain defects. Our research is directed to take LRN instead of Batch normalization (BN), and ADCNN is proposed. BN is used to normalize the data, and then activated by the ReLU function after each convolutional layer AlexNet; then, the maximum pooling operation is performed subsequently. The number of neurons in the last fully connected layer is determined by the number of affective categories. The specific structure of the improved model is shown in Fig. 5. The input setting of the improved model is 227 × 227 × 3, that is, the input is a color image of 227 × 227.

(1) C1 layer: the C1 layer uses 96 convolution kernels of size 11 × 11, and the convolution kernel movement step is set to 4; the size of the output feature map after convolution is 55 × 55 × 96. BN processing is performed immediately after the convolution, and the activation function also uses the ReLU function. When the C1 layer is pooled, the filter size is 3 × 3; the number is equal to the number of feature maps output after convolution, and the filter moving step is set to 2. The size of the output feature map after pooling is 27 × 27 × 96, and the calculation of the size of the feature map after convolution and pooling is shown in formula,

$$size\ (\ feature\_map) \ = \ \frac{\left(size\ \left(input\_data\right)\ +\ 2\ \times\ pad - size\ \left(kernel\right)\right)}{stride} + 1 \quad (17)$$

where, the pad refers to the number of pixels expanded on the feature map, and stride is the moving step of the convolution kernel. In general, when performing pooling operations, the extended feature map is not considered, that is pad = 0.

(2) C2 layer: during the convolution of the C2 layer, the feature map output from the C1 layer is expanded by 2 pixels, and the size of the input feature map becomes 31 × 31 × 96. The size of the convolution kernel used in the C2 layer is 5 × 5, the number is 256, the moving step is 1, and 256 27 × 27 feature maps are output. The C2 layer uses a filter with a size of 3 × 3 and a moving step size of 2 to pool the data, and finally the size of the output feature map is 13 × 13 × 256. Like the C1 layer, the BN and ReLU functions are added between the convolution and pooling of the C2 layer.

(3) C3 layer: Only the data is convolved in the C3 layer, and the convolutional data is not pooled like the C1 and C2 layers. In C3 layer convolution, the feature map output from the C2 layer is first

expanded to 15 × 15 × 256, and then 384 3 × 3 convolution kernels are used to implement the convolution operation with a moving step of 1, and the output is 13 × 13 × 384 feature map. After convolution, the BN and ReLU functions are also used.

(4) C4 layer: The parameter setting during convolution is exactly the same as the C3 layer. The feature map output from the C3 layer is expanded by one pixel and the size is 15 × 15 × 384. After the convolution is completed by 384 3 × 3 convolution kernels, the output 13 × 13 × 384 feature map.

(5) C5 layer: The feature map output from the C4 layer is expanded by one pixel, and the size of the feature map becomes 15 × 15 × 384 during convolution. In the C5 layer convolution, 256 convolution kernels with a size of 3 × 3 and a moving step of 1 are used, and the output feature map size is 13 × 13 × 256. Then, the feature maps activated by the BN and ReLU functions are pooled. When pooling, the filter size is 3 × 3, the moving step size is 2, and the size of the final output feature map is 6 × 6 × 256.

(6) FC6 layer: The FC6 layer uses 4096 neurons to expand the feature map output from the C5 layer into one-dimensional data, Dropout is set to 0.5, and the activation function still uses the ReLU function.

(7) FC7 layer: The FC7 layer is composed of 2048 neurons. This is because the 4096 data of the FC6 layer go through 50% Dropout, and finally output 2048 data. The dropout and activation functions of FC7 layer are the same as FC6 layer.

(8) FC8 layer: The FC8 layer is the classification output layer. The number of neurons is determined according to the number of affective categories. The activation function uses the "softmax" function.

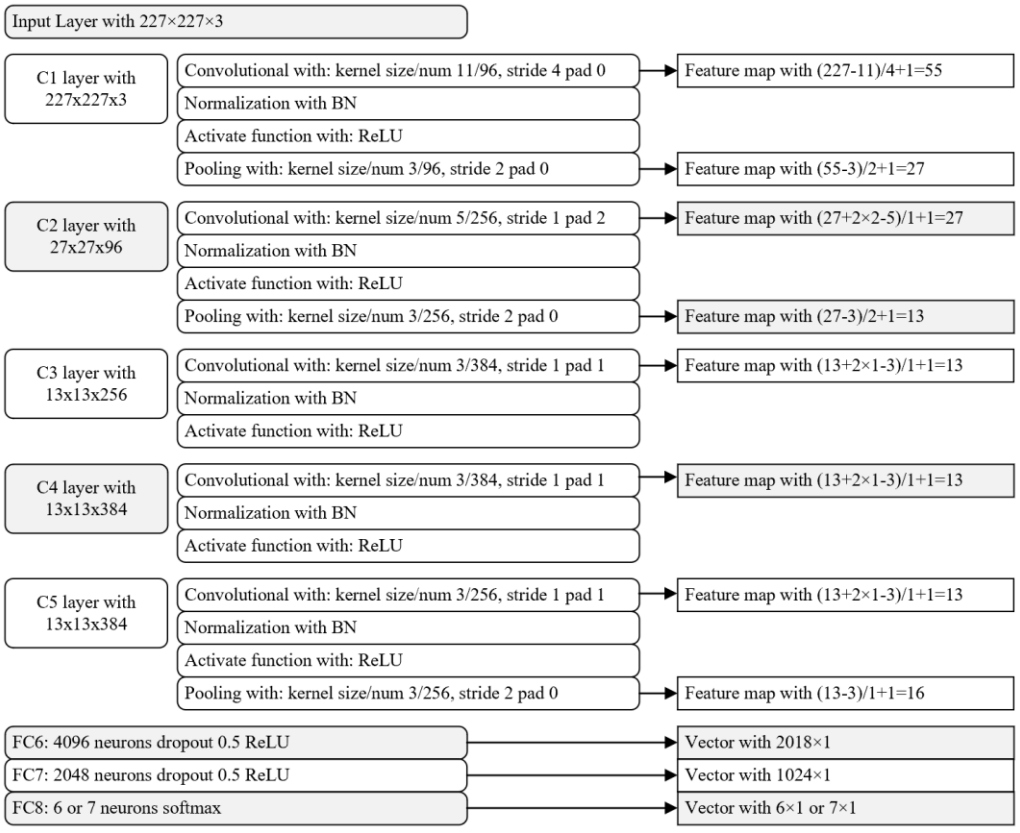


Figure 4. the improved deep neural network by AlexNet

3. Results and discussion

3.1 Data acquisition and preprocessing

The Ryerson audio-visual database of emotional speech and song (RAVDESS, https://smartlaboratory.org/ravdess/) speech audio system was used for training and validating the deep neural networks proposed in this research. The dataset includes speech audio-only files with 16bit, 48kHz, the format is "wav". Full dataset of speech and song, audio and video are approximately 24.8 GB provided by Zenodo. Construction and perceptual validation of 24 actors with 24×60 splits and 8 affective, which named in filename as 01 = "neutral", 02 = "calm", 03 = "happy", 04 = "sad", 05 = "angry", 06 = "fearful", 07 = "disgust", and 08 = "surprised" were prepared; the files also include affective intensity with 01 = normal and 02 = strong [45]. MARSYAS tools were used for calculating partly features that may be used in this research.

3.2 Segmentation

The sound signal segmentation for the waveform information actually takes an amplitude value on the waveform of the analog sound at every time interval, and converts the continuous data in time into the waveform separation process. In the process of extracting affective or affective from sound, proper waveform segmentation can greatly reduce the input complexity of the deep neural networks, thereby reducing the processing time of the deep network can be dramatical improving the accuracy of classification. Here, we show the waveforms of some sounds. According to observations, we know that this segmentation is actually to automatically complete the sound according to the internal structure of the sound signal. As shown in Figure 5. Among them, Figure 5-(a) can be roughly divided into 7 segments, (b) can be divided into 7 segments, (C) is of 6 segments, and (d) is of 7 segments. These divisions are not done under manual intervention, but are formed by extracting key amplitudes.

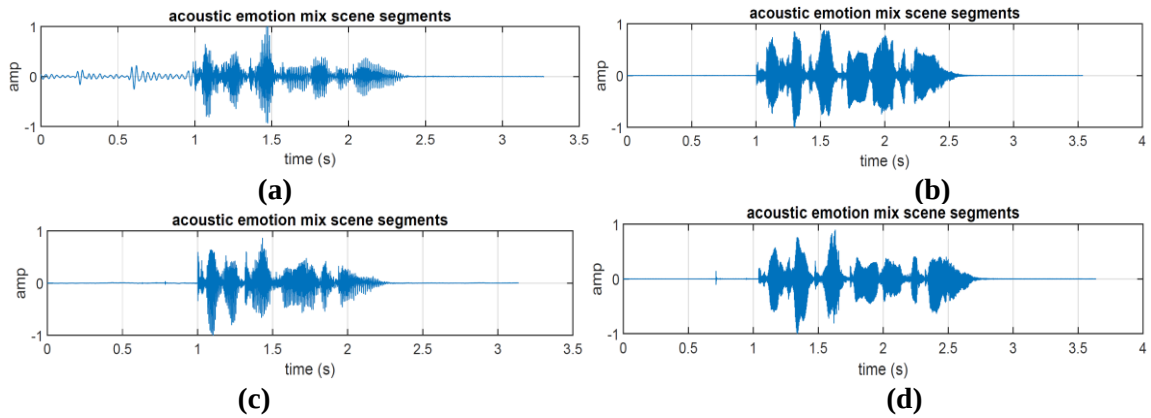


Figure 5. waveform-based segmentation using maximal amplification values

STFT short-time Fourier transform is actually doing FFT on a series of windowed data. Some places will also mention DCT (Discrete Fourier Transform), and the relationship between DCT and FFT is: FFT is a fast algorithm to achieve DCT. Here we set the window length (window.length) to 128; the fast Fourier transform length (FFT.length) to overlap length (overlap.length) to 96. To briefly explain the transformation process, we take Figure 5-(a), and calculated as visualization in Fig.6-(a). There are 12 histogram plots of Mel cep and Fig. 6-(b) is the first histogram for the acoustic signal as showed in Fig. 5(a).

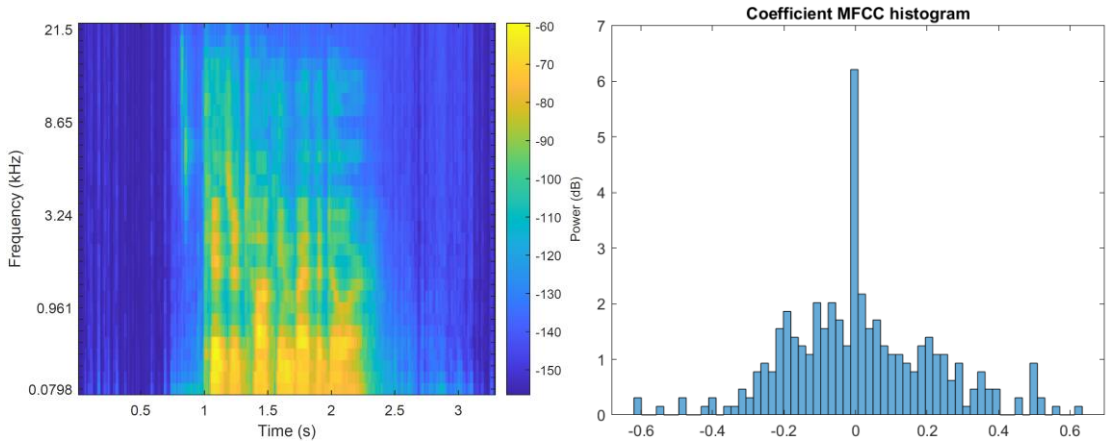


Figure 6. frequency and power of acoustic signal (a) and one histogram plots of MFCC (b)

The formant estimation with linear predictive coding (LPC) coefficients are still calculated subsequently. Fig.7 (a)-(d) show the estimation process of the vowel formant frequencies using LPC. The formant frequencies are obtained by finding the roots of the prediction polynomial. Fig.7 (a)-(d) as relative to Fig.5(a)-(d) use the acoustic samples collected as described in Sec.3.1. the basic feature calculation is based on MARSYAS, while part of calculation works is based on signal processing toolbox™ of Matlab 2019b. The acoustic signal, such as human speech in this case is lowpass-filtered, while the low sampling frequency limits the order of the autoregressive model which can be fitted onto the acquired data. In spite of this limitation, the case here illustrates the technique for using LPC coefficients to determine vowel formants by using segmentation length (segment.length) is 100, and number of overlaps (n.overlap) is 90.

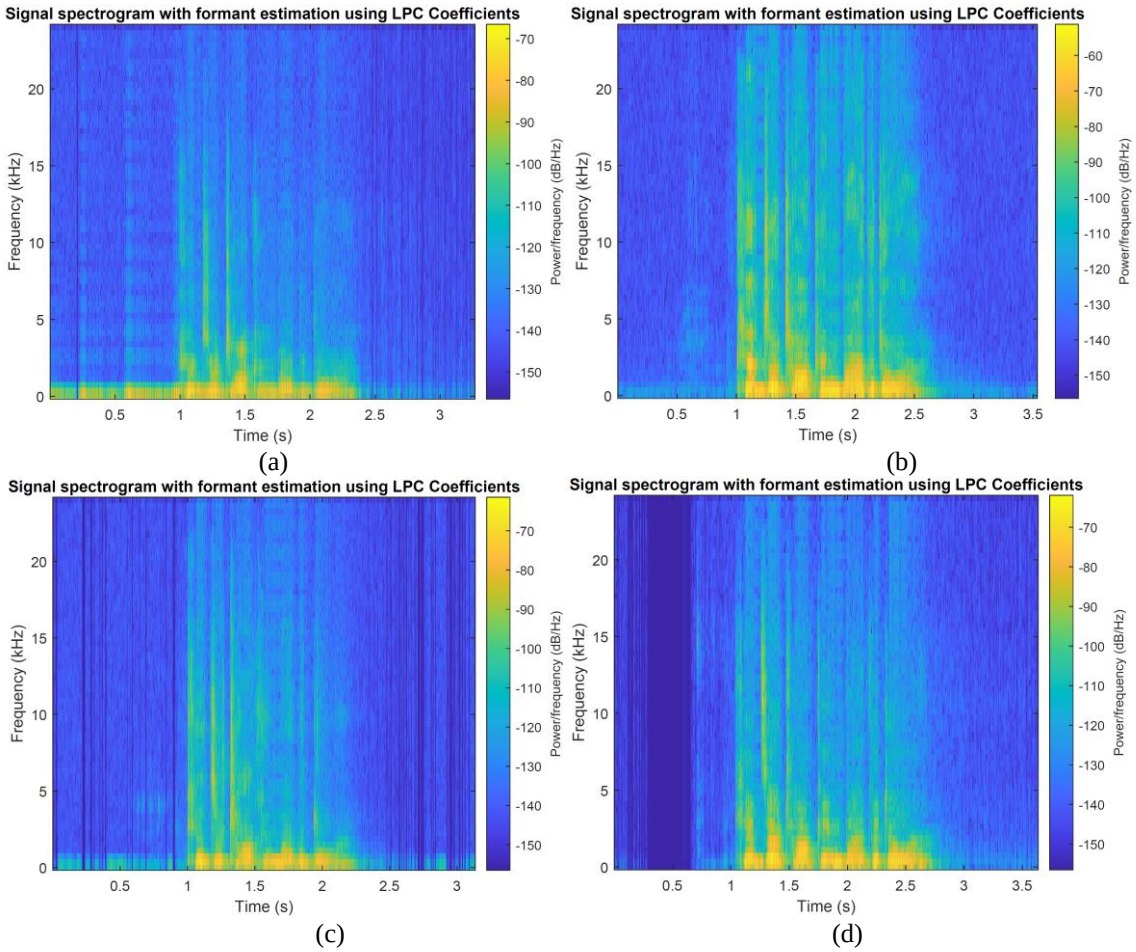


Figure 7. using LPC coefficients to determine vowel formants of acoustic signals

After applied LPC coefficients, the STFT time-frequency representation on segments are calculated as shown in Fig. 8 (a)-(d).

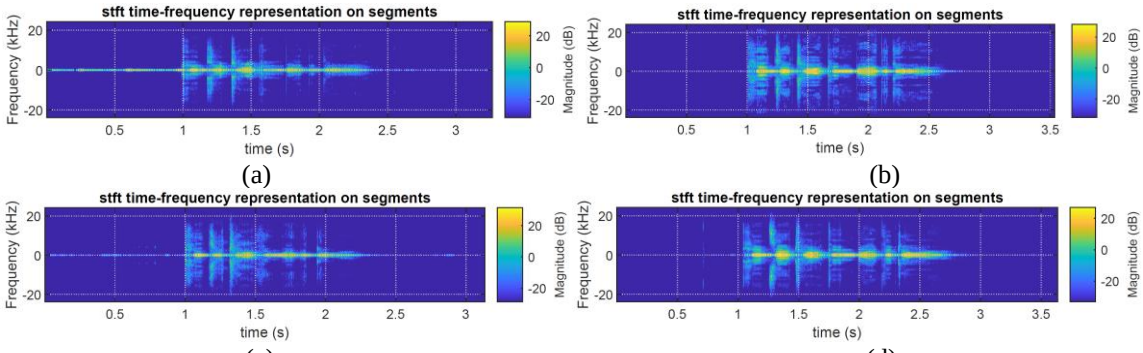


Figure 8. STFT time-frequency presentation on segments of samples (partly)

We also calculated the probability density of the signals with input values, and Fig. 9 shows the concentric distribution of the probability density within [-0.02,0.02], which makes the inputs more stable for the deep neural networks.

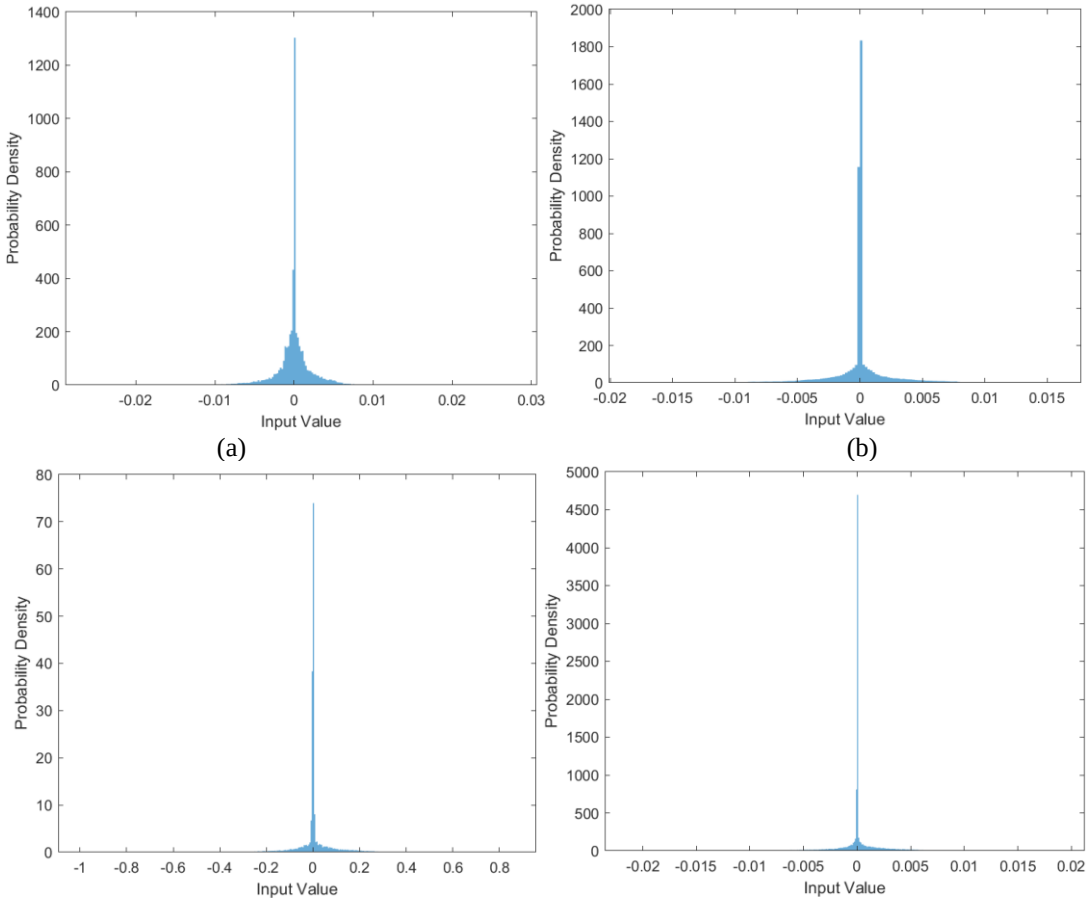


Figure 9. the probability density of the signals with input values

With the SIFT preprocess, the actual improved signals distribution on frequency are shown in Fig.10 (a) - (d), which relative to Fig.5 (a)-(d) separately.

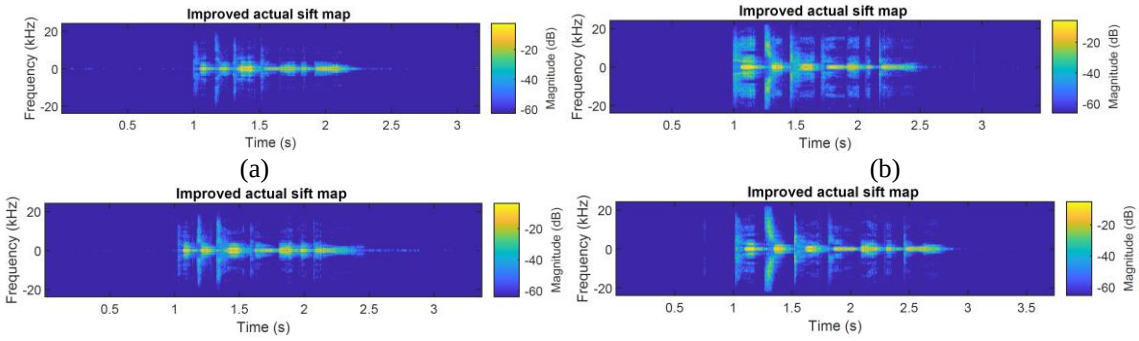
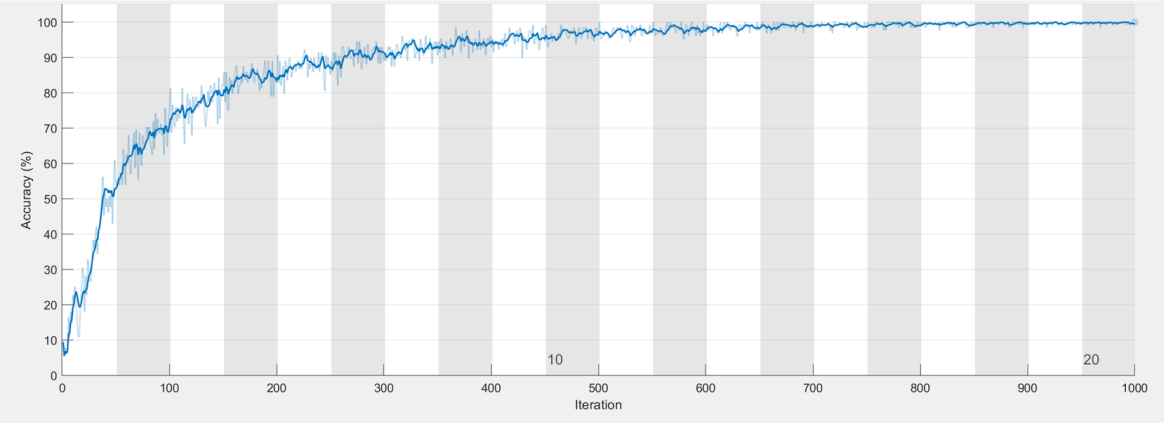


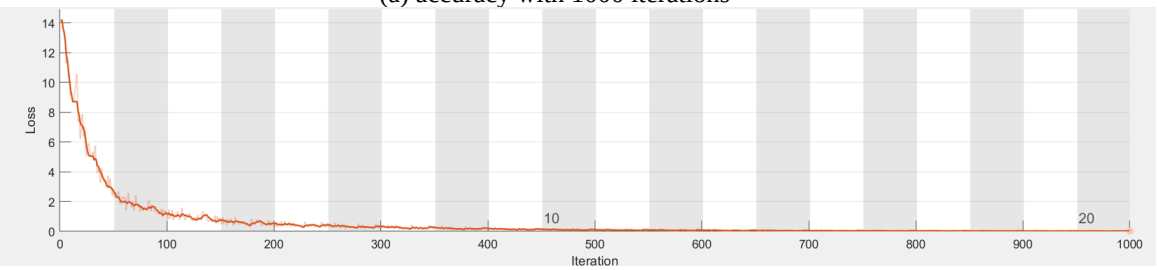
Figure 10. improved actual SIFT distribution of the signals

3.3 Classification using the improved AlexNet based deep convolutional neural network

The proposed improved deep neural networks applied the minimal batch size of 128; tune is 128; momentum is 0.9; L2 regularization is assigned as 0.005; the maximal epoch is 10; piecewise for learning rate with 2 drop period and 0.2 drop factor. In this research, the proposed model used soft-max activation function and finally consumed 182 minutes running time for single CPU (i9-10940 14 cores, 28 threads with 3.5Gb up to 4.2Gb, 32Gb RAM). The accuracy and loss in training the ADCNN model are shown in Fig.11.



(a) accuracy with 1000 iterations



(b) the loss of 1000 iterations

Figure 11. classification accuracy by using a function of iteration in 25 layers of the model (a) and the loss as a function of iteration in 25 layers of the improved AlexNet-based deep neural network.

The kind signal preprocessing and time-frequency transform using SIFT and STFT can improve waveform based acoustic effective classification accuracy. Table 1 shows the results of the comparing results with/without STFT preprocessed signals.

Table 1. proposed method by comparing with/without STFT preprocessed signals

Classifier/ Evaluation metric	**A**	**S1**	**S2**	**PP**	**NP**	**P**	**N**	**F1**
Without SIFT and STFT	93.88%	83.42%	79.69%	81.47%	82.34%	8.18	0.17	0.87
Without SIFT	94.44%	86.23%	81.76%	81.34%	81.23%	8.09	0.18	0.90
Without STFT	95.82%	85.12%	80.23%	81.88%	80.98%	8.10	0.18	0.90
With SIFT and STFT	95.88%	87.98%	82.32%	85.69%	81.56%	8.10	0.18	0.92

Note that, the F1 Score is the 2*((precision*recall)/ (precision + recall)) which conveys the balance between the precision and the recall; it is a normal index for classification evaluation.

From Fig. 12 (a) to (d), it shows that the preprocessed methods have important for the final accuracy, as from Fig. 12-(d), the final accuracy is 95.88% by using the preprocessed method as proposed in Sec. 2.2.1 and 2.2.2. and also, these fusion matrices still show more robustness by inputting SITF and STFT preprocessed signals (the accuracy is from 93.88% to 95.88%).

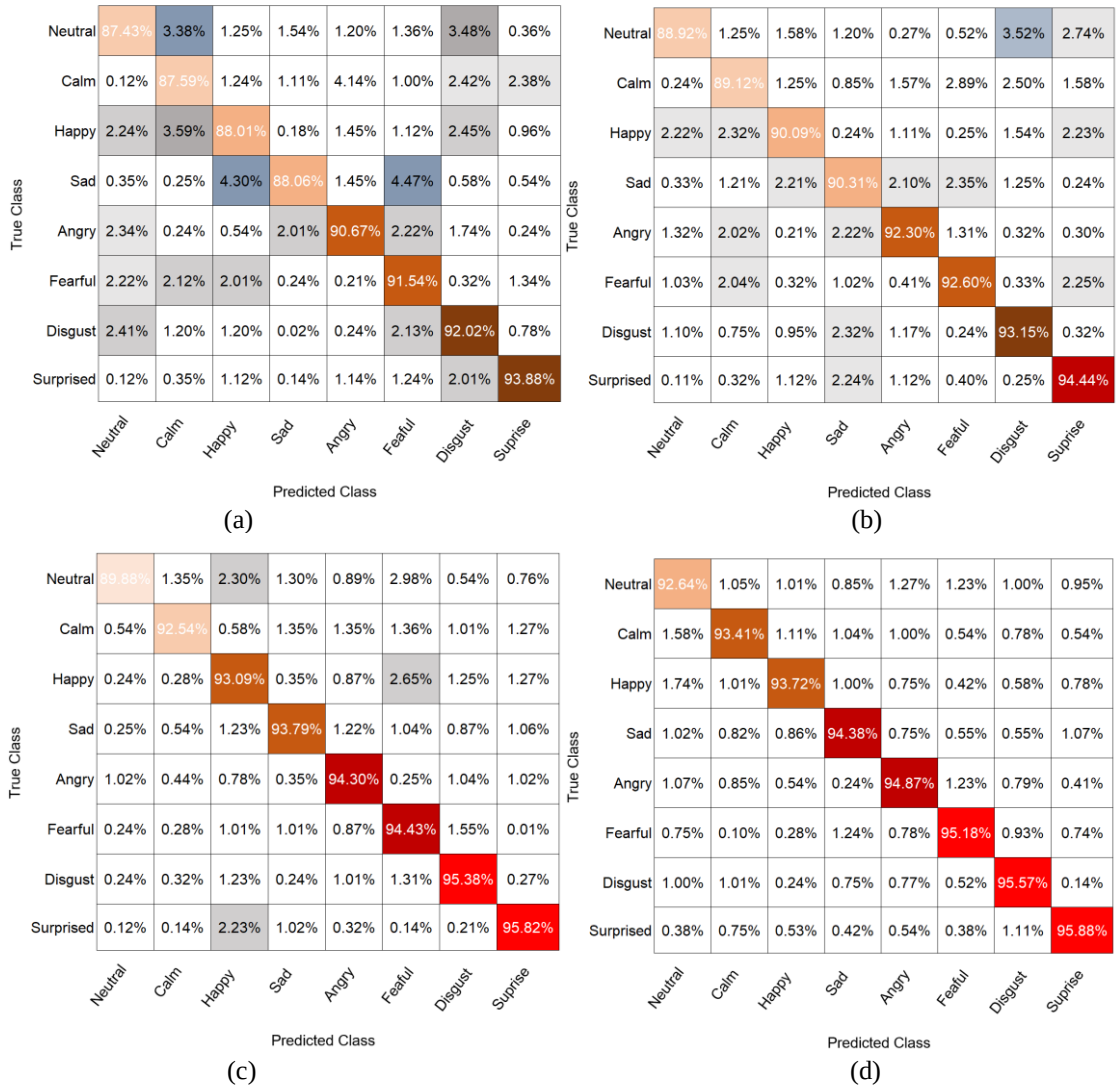
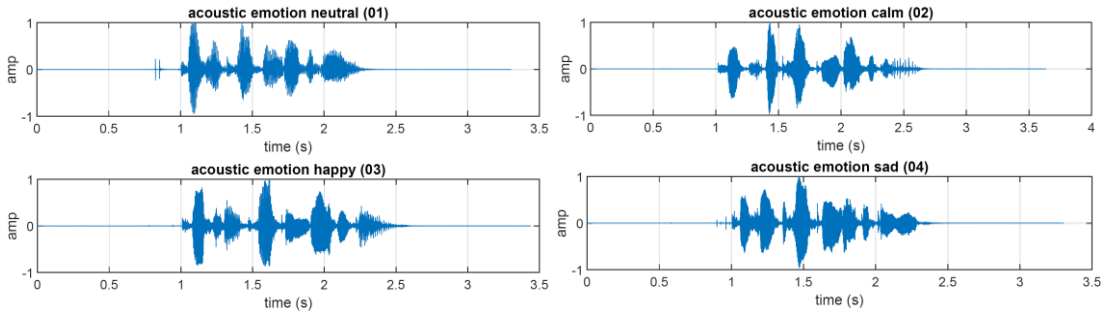


Fig. 12. The confusion matrix of the true class and the predicted class results for eight classification of acoustic signals with (a) the original FFT processed, (b) with MFCC LPC filtered (c) sift maps improved and (d) final results by using the proposed improved AlexNet based deep neural networks.

And the final classification signals are shown in Fig. 13



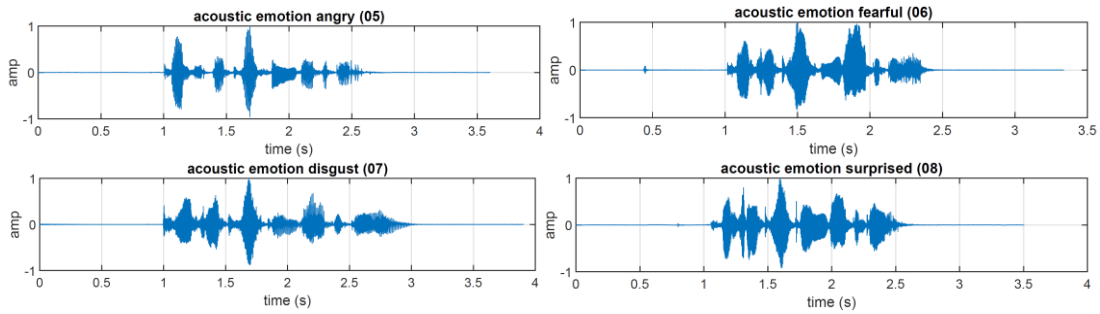


Figure 13. the final affective classified signals

3.4 Discussion

AlexNet based deep neural network was compared with other CNN and combinations of deep neural networks, such as AlexNet+SVM, Pre-trained AlexNet, R-CNN (recurrent convolutional neural network), fast R-CNN, GoogLeNet, VGG-16, scaled conjugate gradient CNN (SGC-CNN), and other linear traditional linear classifiers also were discussed such as decision table (DT), Bayesian inference (BI), artificial neural network (ANN), k-nearest neighbor (kNN), multi-classification support vector MCSVM, long short-term memory (LSTM). To perform the effectiveness of the proposed method, the same acoustic signals were preprocessed by the same algorithms which are introduced in Sec. 2. To deploy comparing analysis, some evaluation indices were adopted, such as, accuracy (A), sensitivity(S1), specific (S2), positive predictive (PP), negative predictive (NP), positive likely (P), negative likely (N), and f1-score (F1). Here we deployed a table for show the performance of the proposed models. Table 2 is for comparing analysis by using the SIFT and /or STFT preprocessed signals. The results show the dominance of the proposed improved AlexNet based deep neural network model in indices of accuracy, S1, PP, N, and F1.

Table 2. Comparative analysis on the different classifiers and the proposed classification model for acoustic signal analysis with STFT preprocessed

Classifier/ Evaluation metric	**A**	**S1**	**S2**	**PP**	**NP**	**P**	**N**	**F1**
DT [46]	84.25%	73.54%	75.56%	82.35%	80.32%	8.21	0.25	0.91
BI [47]	84.27%	85.25%	74.25%	81.52%	78.35%	8.31	0.24	0.91
ANN [48]	84.30%	86.63%	76.36%	80.14%	79.63%	7.56	0.21	0.92
kNN [49]	84.36%	84.25%	78.56%	85.54%	81.25%	7.53	0.15	0.92
MCSVM [50]	84.78%	89.52%	80.25%	84.25%	82.74%	8.35	0.16	0.89
LSTM [51]	86.37%	89.52%	77.58%	84.32%	79.655	8.25	0.20	0.91
R-CNN [52]	87.62%	87.12%	81.25%	85.21%	86.35%	7.36	0.21	0.92
Fast R-CNN [53]	87.64%	87.36%	81.635	85.25%	85.41%	7.89	0.19	0.92
VGG-16 [54]	88.63%	87.35%	82.58%	84.21%	84.82%	8.36	0.20	0.92
SGC-CNN [55]	89.63%	85.45%	82.36%	85.23%	79.85%	8.30	0.18	0.91
GoogLeNet [56,57]	90.67%	87.32%	84.54%	85.30%	82.63%	7.88	0.20	0.91
AlexNet+SVM [58]	90.68%	87.42%	78.69%	85.47%	82.34%	8.14	0.19	0.91
Pretrained AlexNet [59]	93.63%	85.49%	82.56	85.62%	82.98%	8.20	0.18	0.92
ADCNN (*)	**95.88%**	**87.98%**	82.32%	**85.69%**	81.56%	8.10	0.18	**0.92**

4. Conclusions

In general, the artificial neural network with multiple hidden layers is more stable for feature learning. The learned features have a more essential features of the data, which is conducive to visualization or classification; the difficulties to deep neural network may be initial (like layer-wise pre-training) to effectively overcome. In most researches, layer-by-layer initialization is achieved through unsupervised learning and shallow structure algorithm has its limitation lies in the limited ability to express complex functions in the case of limited samples and computing units, and its generalization ability is restricted to a certain extent for complex classification problems. Deep learning can achieve complex function approximation by learning a deep nonlinear network structure, characterize the distributed representation of input data, and demonstrate the powerful ability to learn the essential characteristics of the data set from a few sample sets. The benefit of multi-layer is that it can represent complex functions with fewer parameters. AlexNet carried forward LeNet's ideas and applied the basic principles of CNN to a very deep and wide network; while data preprocessing is successfully applied in this research. By improving the structure and input of the AlexNet based network, and the unique processing method proposed, the classification of acoustic signals has a more accurate result (95.88%).

The shortcomings of the research are that the feature calculation method affects the final recognition efficiency and performance, and the vary features of the acoustic signals and the different calculation processes still may eventually affect the overall accuracy of the recognition. In-depth adjustments to the training process and the network are required to meet actual needs. The algorithms proposed also came up with the following problems, which can be improved in the future, especially the features with the calculation of the number of features and the selection of the number of features; as the number of features increases, the network training speed will increase on a large scale; in addition, the algorithm selection of the preprocessing process needs to be improved in future works. Extracting affective information from acoustic signals is a very difficult process. As the amount of data increases, the accuracy of extraction will also increase, but at the same time, the number of layers of the network need to be considered; while the time consuming of the preprocessing also need to be improved in future.

References

[1] Domínguez-Jiménez, J. A., et al. A machine learning model for affective recognition from physiological signals. Biomedical Signal Processing and Control 2020; 55: 101646.

[2] Imani, M. and G. A. Montazer. A survey of affective recognition methods with emphasis on E-Learning environments." Journal of Network and Computer Applications 2019;147: 102423.

[3] Heysem Kaya, Alexey A. Karpov, Efficient and effective strategies for cross-corpus acoustic emotion recognition, Neurocomputing, 2018; 275: 1028-1034

[4] Kaya, H. and A. A. Karpov. Efficient and effective strategies for cross-corpus acoustic affective recognition. Neurocomputing 2018;275: 1028-1034.

[5] Turner, C. and A. Joseph. A Wavelet Packet and Mel-Frequency Cepstral Coefficients-Based Feature Extraction Method for Speaker Identification. Procedia Computer Science 2015; 61: 416-421.

[6] Wang, K., et al. Wavelet packet analysis for speaker-independent affective recognition. Neurocomputing 2020; 398: 257-264

[7] Boulmaiz, Amira, et al. "Design and implementation of a robust acoustic recognition system for waterbird species using TMS320C6713 DSK." International Journal of Ambient Computing and Intelligence,2017; 8(1): 98-118.

[8] Chia Ai, O., et al. Classification of speech dysfluencies with MFCC and LPCC features. Expert Systems with Applications 2012; 39(2): 2157-2165.

[9] Akçay, M. B. and K. Oğuz. Speech affective recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 2020; 116: 56-76.

[10] Ai, H., et al. DuG: Dual speaker-based acoustic gesture recognition for humanoid robot control. Information Sciences 2019; 504: 84-94.

[11] Zhang, J., et al. Affective recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion 2020; 59: 103-126.

[12] Mannepalli, K., et al. Affective recognition in speech signals using optimization based multi-SVNN classifier. Journal of King Saud University - Computer and Information Sciences.2018; doi:10.1016/j.jksuci.2018.11.012

[13] Özseven, T. A novel feature selection method for speech affective recognition. Applied Acoustics 2019; 146: 320-326.

[14] C.K, Y., et al. Hybrid BBO_PSO and higher order spectral features for affective and stress recognition from natural speech. Applied Soft Computing 2017; 56: 217-232.

[15] Daneshfar, F., et al. Speech affective recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier. Applied Acoustics 2020; 166: 107360.
[16] Wang, X., et al. Human emotion recognition by optimally fusing facial expression and speech feature. Signal Processing: Image Communication 2020; 84: 115831.
[17] Li, X. and M. Akagi. Improving multilingual speech affective recognition by combining acoustic features in a three-layer model. Speech Communication 2019; 110: 1-12.
[18] Sharan, R. V. and T. J. Moir. Acoustic event recognition using cochleagram image and convolutional neural networks. Applied Acoustics 2019; 148: 62-66.
[19] Ali, M. N. Y., Sarowar, M. G., Rahman, M. L., Chaki, J., Dey, N., & Tavares, J. M. R., Adam deep learning with SOM for human sentiment classification. International Journal of Ambient Computing and Intelligence 2019;10(3): 92-116.
[20] Orhan Yaman, Fatih Ertam, Turker Tuncer, Automated Parkinson's disease recognition based on statistical pooling method using acoustic features, Medical Hypotheses 2020; 135:109483
[21] Ke, X., et al. Integrated optimization of underwater acoustic ship-radiated noise recognition based on two-dimensional feature fusion. Applied Acoustics 2019; 159: 107057.
[22] Zhao, J., et al. Speech affective recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control 2019; 47: 312-323.
[23] Yao, Z., et al. Speech affective recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Communication 2020; 120: 11-19.
[24] Tahir, M. A., et al. Training of reduced-rank linear transformations for multi-layer polynomial acoustic features for speech recognition. Speech Communication 2019; 110: 56-63.
[25] S Kumar, VK Solanki, SK Choudhary, A Selamat, RG Crespo, Comparative Study on Ant Colony Optimization (ACO) and K-Means Clustering Approaches for Jobs Scheduling and Energy Optimization Model in Internet of Things (IoT), International Journal of Interactive Multimedia and Artificial Intelligence 2020; 6(1): 107-116
[26] R Gupta, M Khari, V Gupta, E Verdú, X Wu, E Herrera-Viedma, Rubén González Crespo, ``Fast Single Image Haze Removal Method for Inhomogeneous Environment Using Variable Scattering Coefficient, Computer Modeling in Engineering & Sciences 2020; 123(3): 1175-1192
[27] M Khari, AK Garg, R Gonzalez-Crespo, E Verdú, Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks, International Journal of Interactive Multimedia and Artificial Intelligence 2019; 5(7): 22-27.
[28] J Bobadilla, F Ortega, A Gutiérrez, S Alonso, Classification-based Deep Neural Network Architecture for Collaborative Filtering Recommender Systems, International Journal of Interactive Multimedia and Artificial Intelligence 2020; 6(1): 68-77.
[29] Lee, M., et al. Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition. Digital Signal Processing 2019; 85: 1-9.
[30] Papakostas, M. and T. Giannakopoulos. Speech-music discrimination using deep visual feature extractors. Expert Systems with Applications 2018; 114: 334-344.
[31] Ahuja, R., Jain, D., Sachdeva, D., Garg, A., & Rajput, C. Convolutional Neural Network Based American Sign Language Static Hand Gesture Recognition. International Journal of Ambient Computing and Intelligence 2019; 10(3): 60-73.
[32] Chen, W., Li, Y., & Li, C. A Visual Detection Method for Foreign Objects in Power Lines Based on Mask R-CNN. International Journal of Ambient Computing and Intelligence 2020; 11(1): 34-47.
[33] Long, Y., et al. Acoustic data augmentation for Mandarin-English code-switching speech recognition. Applied Acoustics 2020; 161: 107175.
[34] Krizhevsky, A., et al. ImageNet classification with deep convolutional neural networks. J Commun. ACM2017; 60(6): 84-90.
[35] Shanthi, T. and R. S. Sabeenian. Modified Alexnet architecture for classification of diabetic retinopathy images. Computers & Electrical Engineering 2019; 76: 56-64.
[36] Unnikrishnan, A., et al. Deep AlexNet with Reduced Number of Trainable Parameters for Satellite Image Classification. Procedia Computer Science 2018; 143: 931-938.
[37] Boddapati, V., et al. Classifying environmental sounds using image recognition networks. Procedia Computer Science 2017; 112: 2048-2056.
[38] Boloukian, B. and F. Safi-Esfahani. Recognition of words from brain-generated signals of speech-impaired people: Application of autoencoders as a neural Turing machine controller in deep neural networks. Neural Networks 2020; 121: 186-207.
[39] Daldal, N., et al. Automatic determination of digital modulation types with different noises using Convolutional Neural Network based on time-frequency information. Applied Soft Computing 2020; 86: 105834.
[40] Foleis, J. H. and T. F. Tavares. Texture selection for automatic music genre classification. Applied Soft Computing 2020; 89: 106127.
[41] Hou, H.-R., et al. Odor-induced affective recognition based on average frequency band division of EEG signals. Journal of Neuroscience Methods 2020; 334: 108599.

[42] Nicholas Cummins, Shahin Amiriparian, Gerhard Johann Hagerer, et al. An Image-based Deep Spectrum Feature Representation for the Recognition of Affectiveal Speech, MM'17: ACM Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 2017; pp.478-484.
[43] LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. & Jackel, L. D. Backpropagation applied to handwritten zip code recognition. Neural Computation 1989; 1(4):541-551.
[44] Vinod Nair and Geoffrey Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines. //in Proceedings of the 27th International Conference on Machine Learning, ICML 2010, Haifa, Israel, 2010
[45] Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 2018;13(5): e0196391.
[46] Witlox, F., et al. Introducing functional classification theory to land use planning by means of decision tables. Decision Support Systems 2019; 46(4): 875-881.
[47] Li, Z., et al. Brain voxel classification in magnetic resonance images using niche differential evolution-based Bayesian inference of variational mixture of Gaussians. Neurocomputing 2017; 269: 47-57.
[48] Abdel-Hamid, L. Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features. Speech Communication 2020; 122: 19-30.
[49] Priya, T. L., et al. Speech and Non-Speech Identification and Classification using KNN Algorithm. Procedia Engineering 2012; 38: 952-958.
[50] Liu, B., et al. SVM-based multi-state-mapping approach for multi-class classification. Knowledge-Based Systems 2017; 129: 79-96.
[51] Wang Y., Wu Q., Dey N., et al. Deep back propagation–long short-term memory network based upper-limb sEMG signal classification for automated rehabilitation. Biocybernetics and Biomedical Engineering 2020; 40(3):987-1001.
[52] Wang, Y., et al. Classification of mice hepatic granuloma microscopic images based on a deep convolutional neural network. Applied Soft Computing 2019; 74: 40-50.
[53] Wang, D., et al. Optical pressure sensors based plantar image segmenting using an improved fully convolutional network. Optik 2019; 179: 99-114.
[54] Song, Z., et al. Kiwifruit detection in field images using Faster R-CNN with VGG16. IFAC-PapersOnLine 2019; 52(30): 76-81.
[55] F. Shi et al. Texture features based microscopic image classification of liver cellular granuloma using artificial neural networks. 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019, pp. 432-439.
[56] Tang, P., et al. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. Neurocomputing 2017; 225: 188-197.
[57] Esmaeilpour, M., et al. Unsupervised feature learning for environmental sound classification using Weighted Cycle-Consistent Generative Adversarial Network. Applied Soft Computing 2020; 86: 105912.
[58] Shakarami, A., et al. A CAD system for diagnosing Alzheimer's disease using 2D slices and an improved AlexNet-SVM method. Optik 2020; 212: 164237.
[59] Papakostas, M. and T. Giannakopoulos. Speech-music discrimination using deep visual feature extractors. Expert Systems with Applications 2018; 114: 334-344.

Contributors:
Yuxiang Kuang: Conceptualization, Methodology, Data curation, Writing- Original draft preparation. Qun Wu: Visualization, Fuqian Shi and Rubén González Crespo: Investigation, R. Simon Sherratt: Supervision, Ying Wang: Software, Qun Wu and Ying Wang: Validation, Qun Wu, Nilanjan Dey: Writing- Reviewing and Editing.

Authors Bios.

Yuxiang Kuang, conferred his Master's Degree from Beijing Institution of Technology in 2006. His interest in digital art and design led him to pursue a Ph.D. in Zhejiang University. Currently, Dr. Kuang is a Tenured Associate Professor and is also the Director of the Department of Industrial Design in Arts College of Jiangxi University of Finance and Economics. His research topic is the digital art and design, product innovation design, intangible cultural heritage protection. He is the Senior Member of Industrial Design Branch of China Mechanical Engineering Society, the member of China Industrial Design Association, the Expert Committee of Packaging Engineering Magazine.

	**Qun Wu**, is an Associate Professor of Human Factor at the Institute of Universal Design, Zhejiang Sci-Tech University, China. He received his Ph.D. in College of Computer Science and Technology from Zhejiang University, China, in 2008. He holds a B.E. degree in Industrial Design from Nanchang University, China, in 2001, and a M.E. degree in Mechanical Engineering from Shaanxi University of Science and Technology, China, in 2004. His research interests include machine learning, human factor and product innovation design.
	**Ying Wang**, received the B.Eng. degree in Digital Art and Design from Zhejiang University, China in 2014. She is now employed as an Asst. Professor in Department of Industrial Design at College of Art and Design, Zhejiang Sci-tech University, Hangzhou, China. Her research topic is universal design and HCI for elderly.
	**Nilanjan Dey**, was born in Kolkata, India, in 1984. He received his B.Tech. degree in Information Technology from West Bengal University of Technology in 2005, M. Tech. in Information Technology in 2011 from the same University and Ph.D. in digital image processing in 2015 from Jadavpur University, India. In 2011, he was appointed as an Asst. Professor in the Department of Information Technology at JIS College of Engineering, Kalyani, India followed by Bengal College of Engineering College, Durgapur, India in 2014. He is now employed as an Asst. Professor in Department of Information Technology, Techno India College of Technology, India. His research topic is signal processing, machine learning, and information security. Dr. Dey is an Associate Editor of IEEE Access and is currently the Editor-in-Chief of the International Journal of Ambient Computing and Intelligence, and Series co-editor of Springer Tracts of Nature-Inspired Computing (STNIC).
	**Fuqian Shi,** was graduated from College of Computer Science and Technology, Zhejiang University and got his PhD on Engineering, and was a visiting Associate Professor at Department of Industrial Engineering and Management System, University of Central Florida, USA from 2012 to 2014. He is a Senior Member of IEEE, Membership of ACM, and sever as over 30 committee board membership of international conferences; Dr. Shi also serve as associate editors of International Journal of Ambient Computing and Intelligence (IJACI), International Journal of Rough Sets and Data Analysis (IJRSDA), and special issue editor of fuzzy engineering and intelligent transportation in INFORMATION: An International Interdisciplinary Journal. He published over 100 journal papers and conference proceedings, his research interests include fuzzy inference system, artificial neuro networks, and biomechanical engineering.
	**Rubén González Crespo,** Dr. Rubén González Crespo has a PhD in Computer Science Engineering. Currently he is Vice Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is member from different committees at ISO Organization. Finally, He has published more than 200 papers in indexed journals and congresses.
	**R. Simon Sherratt**, is currently Professor of Biomedical Engineering at the University of Reading, UK. Professor Simon Sherratt received the B.Eng. from Sheffield City Polytechnic (now Sheffield Hallam University), M.Sc. from The University of Salford and Ph.D. from The University of Salford; he was elected as Fellow of the IEEE in 2012, Fellow of the IET in 2009; Senior Fellow of the Higher Education Academy in 2014. He is a Chartered Engineer (C.Eng.) and registered European Engineer (Eur Ing). Professor Simon Sherratt was awarded the IEEE International Symposium on Consumer Electronics (ISCE) 2006 1st Place Best paper Award; IEEE Chester Sall Award for best papers in the IEEE Transactions on Consumer Electronics in 2006, 2016, 2017, 2018. He has published over 200 papers in peer review journals and international conferences. His research area is wearable devices, mainly for healthcare and emotion detection.